COMPARAÇÃO DA *LEI DE ZIPF* EM CONTEÚDOS TEXTUAIS E DISCURSOS ORAIS

Comparison of *Zipf's law* in textual content and oral discourse

Rafael-Roeck-Borges Cassettari, Adilson-Luiz Pinto, Rosângela-Schwarz Rodrigues y Letícia-Silvana-dos Santos

Rafael-Roeck-Borges Cassettari, Consultor independente em gestão da informação. Graduado em biblioteconomia pela *Universidade Federal de Santa Catarina* (2014). Especialista em tecnologia de informação e estudos métricos da informação.
http://orcid.org/0000-0002-6633-8253

r.cassettari@grad.ufsc.br

Adilson-Luiz Pinto, professor do *Departamento de Ciência da Informação* da *Universidade Federal de Santa Catarina*. Graduado em biblioteconomia pela *PUC-Campinas* (2000), Mestre em ciência da informação pela *PUC-Campinas* (2004) e doutor em documentação pela *Universidad Carlos III de Madrid* (2007). É Membro do *Grupo de Pesquisa Inteligência, Tecnologia e Informação - Research group* (*ITI-RG*) e líder do *Grupo de Pesquisa InfoCient*. Editor geral da revista *Encontros Bibli*; membro do Conselho editorial das revistas: *Hipertext.net*; *Boletín Millares Carlo*; É especialista em estudos métricos da informação e análise de redes sociais.
http://orcid.org/0000-0002-4142-2061

adilson.pinto@ufsc.br

Rosângela-Schwarz Rodrigues, professora do *Departamento de Ciência da Informação* da *Universidade Federal de Santa Catarina*, é graduada em comunicação social pela *Universidade Federal do Rio Grande do Sul*, mestrado e doutorado em engenharia de produção pela *Universidade Federal de Santa Catarina* (1998 e 2004). É líder do *Grupo de Pesquisa InfoCient*. Membro do Conselho editorial do Portal de periódicos da *UFSC*. Estágio pós-doutoral na *Universitat de Barcelona* junto ao grupo de pesquisa *Cultura i continguts digitals: aspectes documentals, polítics i econòmics* (2012/2013). Atua em acesso aberto e comunicação científica.
http://orcid.org/0000-0002-9639-6390

rosangela.rodrigues@ufsc.br

Letícia-Silvana-dos Santos, mestranda do Programa de pós graduação em ciência da informação da *Universidade Federal de Santa Catarina*, graduada em biblioteconomia pela mesma instituição (2012). Possui especialização em gestão de bibliotecas escolares pela *UFSC*. Faz parte do *Núcleo de Estudos e Pesquisas em Competência Informacional* (*GPCIn*) e do *Grupo de Pesquisa InfoCient*. É Tutora no *Curso de prevenção aos problemas relacionados ao uso de drogas - NUTE/UFSC*. Atua nos estudos métricos da informação e gestão de competências.
http://orcid.org/0000-0002-1447-6590

lety_rugby@hotmail.com

Resumo

A *Lei de Zipf* é uma teoria com base na matemática e na linguística que analisa e quantifica como as palavras são distribuídas dentro de um determinado texto. Desta forma, é possível representar por meio de gráficos e análises estatísticas quais são os termos que mais se repetem, de modo que seja possível criar um ranking de palavras-chave. Esta pesquisa verificou, por meio da *Lei de Zipf*, as variações entre trabalhos acadêmicos escritos e apresentados de forma oral em evento científico. As apresentações orais foram inseridas em forma de vídeo no *YouTube*, para que fosse possível recuperar, de forma auto-

Artigo recibido em 20-11-2014
Aceptação final: 11-02-2015

mática, a transcrição do áudio. Por meio de um script executado em *Bash*, os textos e as apresentações transcritas foram quantificadas e organizadas, sendo possível criar nuvens de tags e tabelas com os rankings, facilitando a comparação entre os conteúdos escrito e oral. Foi possível identificar as esferas dos conteúdos, identificar as palavras em comum ou muito distantes e analisar e comparar matematicamente o que foi escrito com o que foi apresentado oralmente.

Palavras-chave

Lei de Zipf; Bibliometria; Estatísticas linguísticas.

Abstract

Zipf's law is a theory based on mathematics and linguistics that analyzes and quantifies how words are distributed within a text. It is possible to represent by graphs and statistical analyzes which are the terms that are repeated over so that a ranking of keywords is created. This research found, through the *Zipf's law*, variations and uniformities of written academic papers and they presented orally. The oral presentations were inserted in video form on *YouTube*, it was possible to recover automatically the transcript of the audio. Using a Bash script, texts and transcribed presentations were quantified and organized, thereby creating tag clouds and tables with rankings, facilitating the analysis of the contents. It was possible to identify the spheres of content, identifying common words or not and, mathematically, analyze and compare what was written with what was presented in oral discourse.

Keywords

Zipf's law; Bibliometrics; Linguistics statistics.

Cassettari, Rafael-Roeck-Borges; **Pinto, Adilson-Luiz**; **Rodrigues, Rosângela-Schwarz**; **Santos, Letícia-Silvana-dos** (2015). "Comparação da *Lei de Zipf* em conteúdos textuais e discursos orais". *El profesional de la información*, v. 24, n. 2, pp. 157-167.

http://dx.doi.org/10.3145/epi.2015.mar.09

1. Introdução

Hoje, a ciência da informação (CI), apresenta uma série de estudos que envolvem descobertas relevantes para que possa transformar dado em informação e informação em conhecimento; uma série de trabalhos fundamentados na gestão da informação, análise de redes e estudos métricos da informação foram publicados nos últimos anos. Estes últimos, calcados na bibliometria, que tem uma função relevante para a CI, visando analisar a qualidade e quantidade de itens utilizados, consultados, de visibilidade e de produtividade.

A *Lei de Zipf* é uma base matemática-linguística que analisa a frequência e distribuição das palavras contidas em um texto, seja científico ou não. Por meio de cálculo é possível mapear e criar rankings de ocorrência das palavras neste texto (**Zipf**, 1949).

A aplicação desta lei enuncia que existe uma lei do menor esforço (**Egghe**, 1991; **Zörnig**; **Altmann**, 1995; **Günther** *et al.*, 1996), onde algumas palavras/terminologias representam consensos de um campo ou área/subárea, podendo ser considerado trivial as informações de maior relevância, interessante as informações intermediárias e de ruído as informações soltas e de pouca frequência (**Quoniam** *et al.*, 1998).

A *Lei de Zipf* foi explorada para os mais variados idiomas possíveis, como é o caso da aplicação de sua técnica com o idioma chinês e suas representações (**Rousseau**; **Zhang**, 1992; **Shtrikman**, 1994), bem como para a Índia (**Paliwal**; **Bhatnagar**; **Haldar**, 1986), e como para algumas áreas do conhecimento, como a informação nuclear (**Ohnishi**, 1993) e a avaliação linguística (**Baayen**, 1992).

Na área de ciência da informação, iniciativas com a *Lei de Zipf* serviram para simular representações temáticas mais adequadas (**Naranan**; **Balasubrahmanyan**, 1992), inclusive para sustentar aspectos de classificações em unidades de informação e para a probabilidade da recuperação da informação (**Shan**, 2005), com base também na informetria (**Schaer**, 2013).

Mais recentemente, esta lei foi aplicada para as questões semânticas, utilizando a distribuição de frequência, o comportamento de escala temporal e o fator de decaimento (**Zhang** *et al.*, 2008), mesmo que algumas palavras-chave sejam raras nas questões semânticas e pouco aplicadas na questão linguística (**Danesi**, 2009; **Serrano**; **Flammini**; **Menczer**, 2009), praticadas sobre a influência da ciência da computação e da ciência cognitiva (**Chen**; **Chong**, 1992).

Em um contexto mais avançado, temos estudos representando *Zipf* com a projeção visual da informação, como a aplicação de conteúdos *BitTorrent* (BT) e sua escala em localização virtual de textos nos vídeo (**Wang**; **Liu**; **Xu**; 2012), representando a frequência ranqueada; como também o tráfego estatístico da informação no *YouTube* e em sites web 2.0 de compartilhamento de vídeo (**Abhari**; **Soraya**, 2010); a replicação de vídeo de sucesso vislumbrando as palavras-chave transcritas destes conteúdos (**Zhou**; **Xu**, 2007); e inclusive *on-demand* (**Yu** *et al.*, 2006).

Para nossa proposta, foi realizada a aplicação da *Lei de Zipf* em textos escritos e a apresentação oral destes textos, tendo como universo de aplicação três trabalhos acadêmicos apresentados no *14º Encontro nacional de pesquisa em ciência da informação – Enancib*, ocorrido em Florianópolis (Santa Catarina, Brasil). Em um primeiro momento, foram analisados os textos escritos e depois os textos orais. Por fim, foi realizado

o cruzamento e interpretação destes dados a fim de procurar semelhanças ou diferenças nos tipos apresentados.

Desta forma, o estudo objetivou analisar os trabalhos escritos e orais apresentados no *Enancib* de Florianópolis a fim de identificar as semelhanças e diferenças entre as duas alternativas, tendo como especificidade:

- Identificar quais são as três esferas de cada texto (1ª esfera denominada de informação trivial; 2ª esfera denominada de informação interessante e; 3ª esfera denominada de ruído informacional), de modo que seja possível criar um ranking de frequência das palavras encontradas;
- identificar as palavras comuns entre as apresentações orais e escritas;
- comparar a posição de cada termo, para que seja possível analisar as principais diferenças entre o texto escrito e a apresentação oral;
- verificar como as palavras-chave dos textos escritos foram abordadas na apresentações orais;
- comparar e analisar, matematicamente, as diferenças entre os textos escritos e orais.

2. Lei de Zipf

O estudo de *Zipf* trabalha com a frequência de aparição de palavras em um texto, independente do que aquela simbologia represente sozinha ou em conjunto. Por exemplo a palavra "a", em um contexto geral, significa muito mais que apenas a primeira letra de nosso alfabeto mas, nesta lei, ela é apenas mais uma palavra dentro do contexto geral deste texto.

Os autores **Chen** e **Leimkuhler** (1987) nos mostram a importância da *Lei de Zipf* afirmando que ela serve para estudar os mais diversos tipos de fenômenos humanos, montando rankings que exibem, de uma maneira matemática, esses fenômenos que podem ser analisados com mais

Por meio desta base matemática, é possível traçar uma série de fenômenos apenas analisando a frequência em que determinadas palavras aparecem em um texto. A base matemática da lei é:

$$\sqrt{TP} = PE \quad \sqrt{PE} = SE$$

("TP" é o total de palavras - ocorrências; "PE", primeira esfera; "SE", segunda esfera"), ou seja, a raiz quadrada do total de palavras nos dá a primeira esfera; a primeira esfera é, então, palavras que atingem, pelo menos, a raiz quadrada do total de palavras. Por exemplo, se o total de palavras é 625, a primeira esfera são as palavras que tem, pelo menos, 25 (vinte e cinco) ocorrências. A segunda esfera, é a raiz quadrada da primeira esfera, no caso, 5 (cinco). A terceira esfera, que não aparece na fórmula, seria o restante das palavras, que são as que aparecem menos que cinco vezes.

Deste modo, com a *Lei de Zipf*, é possível traçar um ranking de palavras que mais aparecem em um determinado trecho textual ou texto integral. Como destaca **Günther** *et al.* (1996), os dados e o ranking criados pelas informações textuais nada mais são que variáveis aleatórias. É necessário entender o motivo desta ordem de palavras e analisar sua significância perante o estudo.

Ainda no pensamento de **Günther**, entender a razão em que as palavras aparecem na ordem de frequência é o principal motivo da utilização da *Lei de Zipf*. Esta ordem está diretamente ligada com a importância dos termos que aparecem mais vezes dentro do texto.

2.1. Aplicações da *Lei de Zipf*

A *Lei de Zipf* pode ser aplicada de diversas maneiras em análise linguística. A sua aplicação está voltada a textos escritos, onde é realizada uma análise sobre as ocorrências das palavras.

É possível usar a *Lei de Zipf* em estudos de recuperação da informação, como **Quoniam** *et al.* (1998) destacam, analisando as ocorrências em comparação às palavras-chave de determinado texto. Esta aplicação é utilizada para qualificar todas as palavras presentes no texto e, assim, com o seu ranking, entender o assunto do texto ou até comparar se as palavras-chaves selecionadas pelo autor tem realmente alguma ligação como o conteúdo geral do texto.

Segundo um estudo de **Paliwal**, **Bhatnagar** e **Haldar** (1986), a *Lei de Zipf* foi utilizada para traçar a previsão da utilização de determinados recursos, baseando-se na repetição em que cada recurso aparece. Desta forma, é possível mapear e listar estes recursos, focar que material deve ser mais explorado (no caso de minérios, por exemplo), entre outras interpretações.

Outra forma possível de utilização da *Lei de Zipf* poderia ser discursos orais. A diferença é que não há necessariamente um texto escrito. No caso, há uma transcrição do discurso oral para ser feita a análise. A base matemática é a mesma, porém a análise final deve ser um pouco mais cuidadosa, uma vez que o que é dito muitas vezes carrega 'vícios' ou expressões não-formais ou até mesmo regionalistas, que, se não observado corretamente, pode comprometer a análise. Deve-se levar em consideração também as conexões léxicas e suas possíveis variações, como por exemplo singular e plural.

Ridley (1982) escreveu sobre a *Lei de Zipf* focada para os discursos orais. Ele explica que várias palavras irão aparecer muitas vezes, e elas são, por exemplo, "o", "os", "a" e "as", pois são artigos que, em nos idiomas latinos, são frequentemente utilizados em um texto, independentemente de ser extenso ou não e, por outro lado, palavras que são pouco utilizadas, como é o caso de palavras compostas sobre termos específicos como "autodestruição" ou "superestimado". Porém, não devem ser descartadas as diferenças psicológicas em um texto oral e um texto escrito, mesmo que o assunto deles seja o mesmo; os indivíduos tendem a utilizar um vocabulário diferente ao escrever ou falar.

Garner (1997), ressalta que muitas palavras ditas tendem a se repetir mais, mesmo que tenham outros sinônimos. O exemplo dado é a palavra "ready", que em inglês significa "pronto" ou "certo"; esta palavra tem um significado de "entendimento" e que poderia ser facilmente trocada por palavras como "okay" ou "right". Em um texto escrito, optamos sempre por diferenciar as palavras escritas, para dar uma dimensão de vocabulário mais extenso ou até mesmo

em uma questão de estética. Já em apresentações orais, esta questão estética é deixada de lado em termos comuns, como o "ready".

No português isso também ocorre, porém a estrutura e o jeito de escrever pode nos mostrar valores bem diferentes do uso em outras línguas, como o Inglês. Fazer uma análise no idioma português é, em parte uma tarefa mais difícil, por conta das conjugações verbais. A frase "Eu amo você" e "Nós amamos você", aparecem com o verbo "Amar", porém de duas maneiras diferentes. Já no inglês, seria "I love you" e "We love you", onde o verbo que no infinitivo é "to love" aparece duas vezes de maneira igual. Neste caso, a diferença é que, em uma contagem final sobre uma análise fictícia sobre estes textos, no inglês, a palavra apareceria duas vezes e no português, seriam duas palavras diferentes, garantindo a fidelidade com mesmo sentido. Outro exemplo são os artigos definidos que no português são quatro e no inglês apenas um.

3. Metodologia

Este estudo foi baseado em três apresentações, focado na análise do discurso oral de trabalhos apresentados no 14º Enancib e dos mesmos trabalhos em formato textual, fornecido pelo editor científico do evento.

Foram selecionados três trabalhos do grupo de trabalho: produção e comunicação da informação em ciência, tecnologia & inovação, foram gravados oito trabalhos, porém a dificuldade na dicção de alguns trabalhos autores prejudicou as suas seleções. Os três trabalhos selecionados tiveram as falas bem estruturadas e o som nítido para a aplicação desta experiência, e foram definidos como o corpus do estudo.

Os trabalhos selecionados tiveram:

- o primeiro com dois autores, apresentado pelo autor principal;
- o segundo teve três autores, apresentado pela autora principal;
- o terceiro trabalho teve dois autores, apresentado pelo primeiro e principal autor.

Desta forma, toda parte oral teve influência do autor principal de cada um dos trabalhos, para que fosse o mais fiel ao original escrito.

Para o estudo foram filmadas as apresentações e colocadas no *YouTube*, em modo privado, apenas para a coleta das legendas automáticas, onde selecionamos as com mais qualidade em termos de dicção.

Dentro do *YouTube* existe uma ferramenta chamada "Automatic closed captions" ou "Legendas automáticas". Com esta função, é possível que um vídeo em inglês seja transcrito. Nem todas as palavras são interpretadas de maneira correta, há é uma quantidade razoável de erros desta ferramenta para os idiomas latinos, como o caso do português, porém, em inglês, a precisão é altíssima. Desta forma, a coleta de dados foi feita de maneira automática mas, após a geração dos arquivos, foram feitas conferências para correção de erros de forma manual.

O software *YouTube-DL*, é um software, gratuito, escrito na linguagem de programação *Python*. *Python* é uma linguagem de alto nível orientada à objetos. É uma linguagem dinâmica, modular e de multiplataforma. O *Python* é usado de modo geral, pois suas bibliotecas incluem diversas ferramentas desde simples processamentos de texto até utilização de serviços do sistema operacional.

Este software é capaz de baixar integralmente vários dados de um determinado vídeo no *YouTube*. Uma das funções é a possibilidade de baixar as legendas que estão inseridas no vídeo. Com essa função, foi possível fazer a transcrição dos vídeos selecionados e realizar a análise do que foi falado neste vídeo comparando a transcrição do texto da apresentação oral com o texto escrito dos três trabalhos aprovados no evento.

Como complemento foi utilizado um script, executado em *Bash*, para separar e quantificar os dados coletados pela ferramenta *YouTube-DL*. Scripts são um conjunto de instruções de uma determinada linguagem de programação de modo que seja feita uma extensão de sua funcionalidade. Estas instruções são montadas em forma de algoritmo com uma sequência lógica das funções, onde de acordo com as respostas do usuário, o script segue um determinado fluxo do algoritmo.

Bash é um interpretador de comandos em que é possível executar uma série de comandos diretos do sistema operacional. Estes comandos mexem diretamente com operações aritméticas (linguagem das máquinas).
http://www.gnu.org/software/bash

Com os dados coletados (falas em formato texto), foi possível executar o script para quantificar os dados. Nas próprias instruções, os dados gerados foram repassados para um arquivo de extensão .CSV. *comma-separated values* (valores separados por vírgula) é um formato caracterizado como ordenador de bytes. Como o nome sugere, as informações são separadas por vírgula e, utilizando o interpretador certo, os dados foram tabulados automaticamente a partir dos padrões separados pela vírgula.

O comando em *Bash* utilizado foi:

"cat LEGENDAS | tr -cs '[:lower:][:upper:]' '[\n*]' | sort | uniq -c | sed 's/^ *\([0-9]*\) /\1,/' > ARQUIVO.csv"

Este comando executou as seguintes etapas:

- "cat LEGENDAS": "cat" é o comando de leitura de arquivos texto;
- "tr -cs '[:lower:][:upper:]' '[n*]'": "tr" é o comando que copia as entradas principais dos caracteres e repassa para uma nova saída. O parâmetro "-cs '[:lower:][:upper:]' '[n*]'" transforma todos os caracteres de entrada de modo padronizado, inclusive as repetições;
- "sort" - organiza as palavras coletadas;
- "uniq -c" - remove as palavras repetidas e adiciona, ao lado dela, a quantidade de vezes que esta palavra foi repetida no arquivo;
- "sed 's/^ *\([0-9]*\) /\1,/' > ARQUIVO.csv" - cria um arquivo .CSV, com os dados já refinados, em ordem alfanumérica do modo "Palavra, número de repetições".

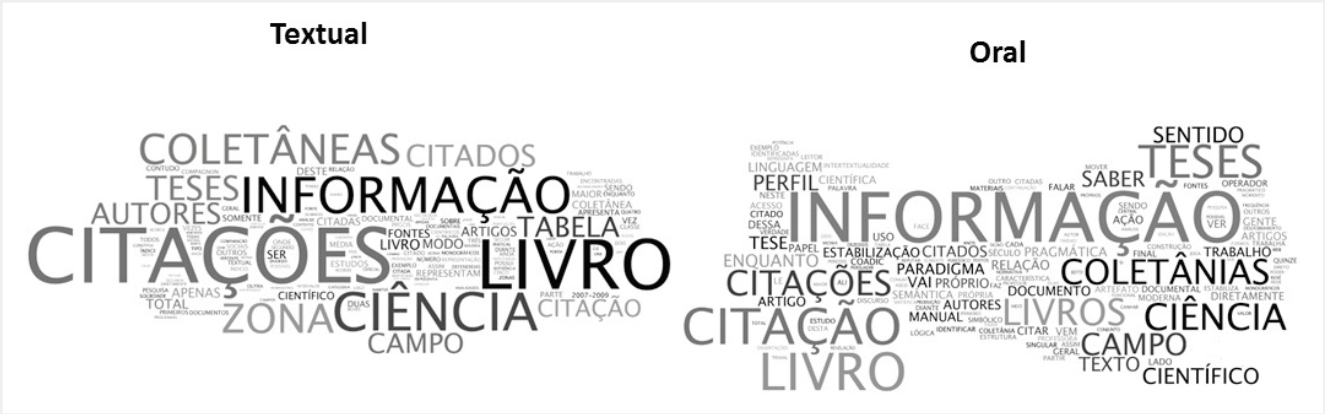


Figura 1. Nuvem de palavras do conteúdo 1 (textual e oral)

Como a língua portuguesa possui caracteres especiais (acentuações), utilizamos outro comando para a remoção dos acentos. Entretanto algumas palavras perdem o sentido por conta da sua acentuação, porém pouco interferiu na somatória ao ponto de interferir no resultado das amostras.

O comando utilizado previamente nos arquivos de texto foi:

"sed'y/áÁàÀãÃâÂéÉêÊíÍóÓõÕôÔúÚçÇ/aAaAaAaAeEeEiIoOoOoOuUcC/' ARQUIVO_DE_ENTRADA.txt > ARQUIVO_DE_SAÍDA.txt"

A função deste comando serviu para substituir todos os acentos (primeira parte do comando) pelas letras limpas (segunda parte do comando).

Por outro lado, em nossa aplicação, tratamos estes dados nas esferas da *Lei de Zipf*, tendo como:

- Primeira Esfera: palavras mais comuns dentro do texto. São incluídos artigos definidos e indefinidos, preposições, conectores, etc.
- Segunda Esfera: palavras mais importantes dentro do texto. São as palavras que ocorreram em menor número que as palavras citadas da primeira esfera. Por não serem palavras de uso comum, serão consideradas como importantes;
- Terceira Esfera: palavras consideradas como ruídos dentro do texto.

Com o auxílio de um software que gerencie planilhas, como o 'BrOffice Calc', foi possível analisar os dados coletados e fazer comparações entre o discurso e o conteúdo textual.

Todo o tratamento teve preocupação exclusiva de analisar conteúdos textuais individuais (palavras individuais), porém o mesmo poderia ser aplicado a palavras compostas, entretanto para esta realização teríamos que controlar as informações de forma individual, fato que tornaria o trabalho mais custoso e pouco funcional em se tratando de um método automatizado, como o proposto desde o inicio.

Outro detalhe da análise foi que não normalizamos os termos/palavras, entretanto o mesmo poderia ser realizado segundo singular/plural e sinônimos.

4. Resultados e discussão

As análises foram feitas de maneira individual. Serão apresentadas as análises em todos os conteúdos textuais e orais, de maneira que seja possível interpretar as correlações entre ambos pela *Lei de Zipf*.

O primeiro conteúdo analisado tem como título *Os livros nas teses da ciência da informação: um estudo de citação*, contendo dois autores, com a apresentação oral do primeiro autor.

Este trabalho teve um total de 1.226 palavras diferentes em um total de 4.379 palavras escritas. A palavra com o maior número de ocorrências foi 'a' com 202 repetições. Aplicando a fórmula de *Zipf*, encontramos as três esferas textuais divididas nos seguintes valores:

- primeira esfera: 48 palavras com até 14 repetições;
- segunda esfera: 213 palavras com até 3 repetições;
- terceira esfera: 966 palavras.

Na primeira esfera do conteúdo textual é importante destacar as palavras "citação" e "citações" que aparecem 81 e 23 vezes respectivamente somando 104 e que a palavra "ciência" e "informação" aparecem 56 e 41 vezes respectivamente. É possível perceber a importância que estas palavras têm dentro do seu escopo, uma vez que estes termos se encontram, inclusive, no título do trabalho.

Em sua segunda esfera, as palavras que tem o maior destaque são "artigos", "científico", "fontes" e "informação". Também em uma análise partindo do título do texto, a importância destas palavras também devem ser destacadas, uma vez que elas são os termos que puxam o início da segunda esfera, que é considerada interessante, pois pode ser reutilizada em um segundo momento, caso seja necessário para o entendimento do contexto informacional.

Para a análise oral o conteúdo teve 571 palavras únicas e 2.064 repetições. A palavra com o maior número de ocorrências foi "e" com 122 aparições. Suas esferas ficaram divididas em:

- primeira: 32 palavras com até 11 repetições;
- segunda: 107 palavras com até 3 repetições,;
- terceira: 432 palavras.

A primeira esfera teve 32 palavras onde as oito primeiras são artigos/preposições, apenas na nona posição aparece a palavra "informação", com 32 ocorrências. Logo em seguida aparecem as palavras "livro", "citação", "teses", "ciência", "livros" e "citações". Todas estas palavras mencionadas aparecem e/ou tem sentido direto com o título do texto apre-

Figura 2. Nuvem de palavras do conteúdo 2 (textual e oral)

sentado. Se considerar as palavras "livro" e "livros" como uma só, elas somam 41 ocorrências, elevando-as à oitava posição e, no mesmo caso, com as palavras "citação" e "citações", juntas somariam 39 ocorrências, passando para a nona posição. Outra palavra que deve ser destacada é a palavra "coletânea" que junta 17 ocorrências.

Na segunda parcela, 107 protagonizam o que chamaríamos de parte interessante onde as palavras que dominam este topo são "científico", "perfil", "saber" e "sentido", todas com 10 ocorrências. A palavra "tese" aparece em 8 oportunidades e que, se somada com a palavra "teses" que apareceu na primeira esfera, somariam 29 e ocupariam a décima posição. Outras palavras interessantes que possuem grande relação com o texto são "pragmática", "relação", "documento", "autores", "artigo", com 6 ocorrências e "científica" e "semântica", com 5 repetições.

O segundo conteúdo estudado foi o trabalho Redes de coautoria em ciência da informação: estudo sobre a natureza das relações na área temática de organização e representação do conhecimento de três autores, apresentado pela primeira autora.

A parte textual conteve 1.190 palavras diferentes representadas em total de 4.753 repetições. O maior número de ocorrências ficou com a palavra "de" com 311 frequências. A divisão das esferas ficou:

- primeira: 52 palavras com até 17 repetições;
- segunda: 123 palavras com até 4 repetições;
- terceira: 1009 palavras.

Sua primeira esfera, teve a quinta palavra e a primeira que não são artigo/preposição/conector é "relações" com 84 repetições e a palavra "relação" aparece 53 vezes. Se considerar elas como uma palavra só, excluindo o plural, ela se repete 137 vezes, o que deixaria ela na quarta posição com uma proporção de 5,49% da primeira esfera. Tendo em vista que as únicas palavras que passariam ela são "a", "e" e "de", a palavra relação/relações tem uma importante função dentro do texto.

Outra análise interessante, ainda dentro da primeira esfera é que apenas as palavras "redes" e "relação/relações" são palavras que se encontram no título.

Na segunda esfera, aparecem as palavras "ciência", "informação" e "área", que são palavras do título, de maneira expressiva. As palavras com o maior número de aparições são "menor", "incidência" e "foi" que, de maneira individual, não teriam muito sentido. Outras palavras expressivas na segunda esfera foram "pesquisadores" com 13 ocorrências, "orientandos" com 11, "especialização" com 9, "orientação" com 9 e "pesquisador" com 9. Neste caso, se considerar as palavras "pesquisador" e "pesquisadores", sem o plural, elas atingem um total de 22 ocorrências, que promoveriam elas à primeira esfera.

O conteúdo oral apresentou um total de 660 palavras e 2809 ocorrências e a palavra com mais ocorrências foi "de" em 188 oportunidades. As esferas foram divididas no seguinte modo:

- primeira: 42 palavras com até 13 repetições;
- segunda: 134 palavras com até 3 repetições;
- terceira: 484 palavras.

Na primeira divisão há uma palavra que merece uma atenção especial por se tratar de um prefixo. A palavra "co" aparece 101 vezes sendo que, analisando o texto separadamente, percebe-se que ela aparece sempre em duas oportunidades. Ou aparece como "co-citação(ões)" ou como "co-autores/autoria". A palavra "citação" aparece 68 vezes e a palavra "citações" aparece 14 vezes, dando um total de 82 repetições. A palavra "autores" aparece 56 vezes e a palavra "autoria" 14 vezes, onde a soma é 70. Analisando em uma maneira geral, sem qualquer análise individual no texto, a soma dessas ocorrências é de 152 aparições e que 101 vezes elas aparecem juntas do prefixo "co", que tem grande ligação com o título e contexto do trabalho apresentado.

Ainda na primeira esfera, há outras palavras que devem ser destacadas como "citados", que aparece 26 vezes, "frequência", 20 vezes, "Salton", 13 vezes. A palavra "Salton" aparece na primeira esfera por se tratar no "Cosseno de Salton" onde, inclusive, a palavra "cosseno" aparece na segunda esfera com 10 repetições.

O último conteúdo analisado é intitulado *Webometria e produção conjunta entre universidades federais da região sul do Brasil*, de dois autores, apresentado pelo primeiro autor.

A produção textual obteve um total de 928 palavras únicas e 4834 palavras escritas. A palavra que mais se repetiu foi 'de' com 297 ocorrências. Os grupos ficaram divididos da seguinte forma: (primeira esfera) 48 palavras com até 17 repetições; (segunda esfera) 166 palavras com até 4 repetições, e; (terceira esfera) 715 palavras.

Figura 3. Nuvem de palavras do conteúdo 3 (textual e oral)

Este texto foi o que trouxe a maior quantidade de palavras realmente importantes na primeira esfera. As palavras "páginas", "citações", "universidades", "conexões", "UFRGS", "UFSC", "websites", "web", "universidade", "UFPR", "instituições", "trabalhos", "UFSM", "operadores", são exemplos de palavras que contém um importante sentido ao texto e aparecem logo na primeira esfera. Em uma análise direta com o título, a palavra "webometria" acaba não aparecendo, porém a relação entre universidades federais da região sul do Brasil e as siglas da universidade já nos mostra a relação entre o título e as repetições. Outra informação de grande destaque é a palavra "páginas", que aparece em 82 oportunidades e "websites", 39 vezes, podem ser consideradas sinônimos e, desde modo, juntas garantem a soma de 121 aparições, o que levaria elas à quinta posição.

Na segunda esfera é possível encontrar alguns termos muito interessantes e relevantes ao trabalho. Palavras como "[illegible]" e "[illegible]" refletem a importância dos operadores, citados 21 vezes na primeira esfera, no trabalho apresentado. Outras palavras interessantes são "webométricos", "*WoK*" (*Web of Knowledge*), "*Google*", "produção" e "produções". As palavras "produção" e "produções", se somadas, aparecem 27 vezes, de modo que fossem incluídas na primeira esfera.

Para a parte oral o conteúdo teve um total de 577 palavras únicas e 2.326 repetições. A palavra que mais se repetiu foi "e" com 128 vezes. A divisão das esferas foram:

- primeira esfera: 44 palavras com até 11 repetições;
- segunda esfera: 133 palavras com até 3 repetições;
- terceira esfera: 400 palavras.

A primeira esfera tem como principais termos as palavras "universidades" e "universidade", onde tem 31 e 28 repetições respectivamente. Analisando de maneira coletiva as palavras, juntas somariam 59 repetições, que às levariam para a quinta posição, perdendo apenas para os termos "e", "que", "a", e "de". O termo seguinte é "UFRGS", com 24 aparições na décima sexta posição, o termo "'página", com 19, "UFSC", com 17. Em uma forma coletiva, as palavras "website", "websites" e "página" somariam 47 ocorrências, colocando-as na sexta posição, sendo o segundo termo mais importante da apresentação oral.

Sua segunda esfera apresenta como termo principal a palavra "título", com 10 aparições. Ainda nesta esfera, é possível encontrar as palavras "âncora" e "URL", com 9 e 7 repetições respectivamente. Estas palavras, como dito na análise do texto escrito, são operadores utilizados na elaboração do trabalho. Outros termos interessantes encontrados nesta esfera são "*UFPR*", "*UFSM*", "*FURG*" e "*UFPEL*", que são as universidades analisadas.

A palavra "webometria", que se encontra no título do trabalho foi dita apenas 1 vez durante a apresentação: na leitura do título.

Uma análise interessante é que, nas considerações do conteúdo textual, as universidades de maior importância destacadas são *UFRGS* e *UFSC* e que são, também, as mais citadas na apresentação oral.

4.1 Análise comparativo dos conteúdos textuais X orais

Foi estudada a composição comparativa dos 3 conteúdos (textos e apresentações), onde contextualizamos a influência das palavras em identificação de conteúdos e no ranqueamento em que aparecem as palavras.

Para uma melhor visualização iremos dividir cada um dos 3 conteúdos em blocos.

Bloco do conteúdo 1

Esta primeira tabela serve para termos uma ideia dos valores coletados de uma maneira individual, para em um segundo momento levantar algumas conclusões.

É possível, por meio desta tabela, ver a relação entre as palavras escritas e as palavras ditas. Alguns resultados, como:

- a relação Escrito/Oral das palavras únicas é de 2,14:1 e de palavras totais é de 2,12:1;
- a média entre as palavras totais e as palavras únicas no texto escrito é de 3,57% palavras;
- a média entre as palavras totais e as palavras únicas no texto oral é de 3,61% palavras;
- a primeira esfera do trabalho escrito representa 3,91% das palavras únicas;
- a primeira esfera do trabalho oral representa 5,6% das palavras únicas;
- a segunda esfera do trabalho escrito representa 17,37% das palavras únicas;

Tabela 1. Tabela de dados extraídos do conteúdo 1

Identificação de conteúdos	Textual	Oral
Palavras únicas	1.226	571
Total de palavras	4.379	2.064
Palavras na primeira esfera	48	32
Repetições dentro da primeira esfera	14	11
Palavras na segunda esfera	213	107
Repetições dentro da segunda esfera	13 entre 3	10 entre 3

- a segunda esfera do trabalho oral representa 18,73% das palavras únicas.

Neste primeiro conteúdo é possível perceber a semelhança proporcional entre a apresentação oral e o conteúdo textual.

A tabela 2 mostra as 10 primeiras palavras da primeira esfera do texto escrito junto com o seu Ranking e aparição nos textos orais e escritos. Nesta tabela foram excluídos os artigos, preposições e conectores, porém foi mantida a posição do ranking. O critério de desempate utilizado é a organização alfanumérica.

As palavras com maior ocorrência no texto escrito tem uma posição inversa no texto oral; A palavra "informação" é a que tem a melhor posição no texto oral, ocupando a 9ª posição, mas, no texto escrito, é apenas a quinta mais importante, ocupando a 20ª posição; A palavra "citação" obteve a maior posição no texto escrito, ocupando a 6ª posição e no texto oral ocupa apenas a 25ª posição; Entre as 10 primeiras palavras da primeira esfera do texto escrito, apenas 7 também se encontram na primeira esfera do texto oral; A palavra "Zona" é a sexta palavra com mais aparições no texto escrito, ocupando a 22ª posição, porém ela não foi dita uma única vez no texto oral. A palavra "zonas" foi dita apenas uma vez.

Com complemento temos algumas palavras na tabela 2 que priorizam o conteúdo oral, como:

- as palavras "livro", "campo" e "científico" não aparecem no ranking das 10 palavras que mais se repetem no texto escrito;
- a palavra 'campo' tem uma importância semelhante tanto no texto escrito como no texto oral;
- apenas a palavra 'científico' não está na primeira esfera do texto escrito.

Bloco do conteúdo 2

Com base no conteúdo, é possível destacar que:

- a relação escrito/oral das palavras únicas é de 1,8:1 e de palavras totais é de 1,69:1;
- a média entre as palavras totais e as palavras únicas no texto escrito é de 3,99% palavras;
- a média entre as palavras totais e as palavras únicas no texto oral é de 4,25% palavras;
- a primeira esfera do trabalho escrito representa 4,36% das palavras únicas;
- a primeira esfera do trabalho oral representa 6,3% das palavras únicas;

Tabela 2. Ranking de palavras do conteúdo 1

Palavras	Número de aparição textual	Posição no ranking textual	Número de aparição oral	Posição no ranking oral
Citações	81	6ª	16	25ª
Livros	73	8ª	17	23ª
Coletâneas	44	17ª	17	21ª
Ciência	41	19ª	17	20ª
Informação	41	20ª	32	9ª
Zona	37	22ª	-	-
Teses	36	24ª	21	15ª
Citados	32	27ª	7	48ª (2ª esfera)
Autores	31	28ª	6	53ª (2ª esfera)
Citação	23	34ª	23	12ª
Campo	29	30ª	14	29ª
Livro	16	41ª	24	11ª
Científico	12	51ª (2ª esfera)	10	30ª

- a segunda esfera do trabalho escrito representa 10,33% das palavras únicas, e;
- a segunda esfera do trabalho oral representa 20,3% das palavras únicas.

Como no texto anterior, este trabalho tem uma semelhança grande entre a relação de palavras únicas e totais em ambos formatos de apresentação. A proximidade entre as proporções da primeira esfera também é visível.

A tabela 4 mostra as primeiras palavras mais relevantes da primeira esfera dos conteúdos, juntamente com o seu ranking e aparição nos textos orais e escritos do segundo trabalho analisado.

As informações que podemos retirar da análise desta tabela é que:

- do mesmo modo que o trabalho analisado anteriormente, as posições são inversas no texto escrito e oral;
- não houve nenhuma ocorrência em que as 10 primeiras palavras da primeira esfera no texto escrito estivessem no texto oral;
- 5 termos que mais se repetiram no trabalho escrito não tiveram qualquer aparição na apresentação oral;
- não há palavras da primeira esfera oral na primeira esfera escrita;
- duas palavras não foram mencionadas no texto escrito e outras quatro foram apenas mencionadas uma única vez;
- do mesmo modo que a análise do texto escrito, o texto oral utiliza outras palavras para ser apresentado;

Tabela 3. Tabela de dados extraídos do conteúdo 2

Identificação de conteúdos	Textual	Oral
Palavras únicas	1.190	660
Total de palavras	4.753	2.809
Palavras na primeira esfera	52	42
Repetições dentro da primeira esfera	17	13
Palavras na segunda esfera	123	134
Repetições dentro da segunda esfera	16 entre 4	12 entre 3

Tabela 4. Ranking de palavras do conteúdo 2

Palavras	Número de aparições textuais	Posição no ranking textual	Número de aparições oral	Posição no ranking oral
Relações	84	5ª	2	270ª (3ª esfera)
Relação	53	11ª	5	102ª (2ª esfera)
Redes	48	14ª	3	163ª (2ª esfera)
Pesquisa	45	16ª	3	151ª (2ª esfera)
Orientador	39	19ª	-	-
Científica	36	20ª	6	79ª (2ª esfera)
Nível	36	21ª	-	-
Orientando	36	22ª	-	-
Mestrado	35	24ª	-	-
Doutorado	30	28ª	-	-
Autoria	9	75ª (2ª esfera)	14	33ª
Análise	8	87ª (2ª esfera)	38	11ª
Autores	3	191ª (2ª esfera)	56	8ª
Frequência	3	218ª (2ª esfera)	20	22ª
Citação	1	593ª (3ª esfera)	68	6ª
Citações	1	594ª(3ª esfera)	14	34ª
Domínio	1	716ª (3ª esfera)	18	26ª
Proximidade	1	1051ª (3ª esfera)	17	30ª
Matriz	-	-	15	32ª
Citados	-	-	26	17ª

- com base apenas nestes dados, é possível afirmar que o texto oral e escrito não estão alinhados;
- como o trabalho escrito foi criado antes, pode-se dizer que a apresentação oral não condiz com o que foi apresentado na forma escrita.

Bloco do conteúdo 3

Com as informações da tabela 5 é possível dizer que:

- a relação escrito/oral das palavras únicas é de 2,07:1 e de palavras totais é de 1,6:1;
- a média entre as palavras totais e as palavras únicas no texto escrito é de 5,2% palavras;
- a média entre as palavras totais e as palavras únicas no texto oral é de 4,03% palavras;
- a primeira esfera do trabalho escrito representa 5,17% das palavras únicas;
- a primeira esfera do trabalho oral representa 7,62% das palavras únicas;
- a segunda esfera do trabalho escrito representa 17,88% das palavras únicas, e;
- a segunda esfera do trabalho oral representa 23,05% das palavras únicas.

Tabela 5. Tabela de dados extraídos do conteúdo 3

Identificação de conteúdos	Textual	Oral
Palavras únicas	928	577
Total de palavras	4.834	2.326
Palavras na primeira esfera	48	44
Repetições dentro da primeira esfera	17	11
Palavras na segunda esfera	166	133
Repetições dentro da segunda esfera	16 entre 4	10 entre 3

Utilizando como base os textos anteriores, o terceiro texto segue alguns dados que podemos afirmar que acabam se padronizando em uma apresentação oral de um trabalho escrito, onde a relação de palavras são sempre valores bem próximos (de modo proporcional).

A tabela 6 traz o ranking das palavras que mais se repetem no terceiro conteúdo, onde a apresentação oral foi a única que trouxe todas as 10 primeiras palavras da primeira esfera do trabalho escrito, onde 2 palavras se encontram na segunda esfera e 1 na terceira esfera do texto oral.

O posicionamento (ranking) das palavras da primeira esfera do texto escrito são, relativamente, parecidas na apresentação oral, destacando as palavras "universidades", "websites", "*UFRGS*" e "*UFSC*"; diferentemente dos outros conteúdos, não demonstrou uma relação inversa no posicionamento dos termos.

De acordo com os dados, podemos afirmar que:

- as principais palavras do discurso oral fazem parte do texto escrito onde em 6 oportunidades aparecem na primeira esfera dos formatos apresentados;
- a palavra com menos repetições relativas no texto escrito é "website", com 10 repetições, porém, se levar em consideração que o seu plural "websites" atingiu a marca de 39 repetições, elevaria o termo a um dos mais utilizados onde apareceria por 49 oportunidades.

O cenário semelhante detalhado anteriormente também ocorre com a palavra "página" em que, no trabalho escrito, não se repete tantas vezes mas no plural é a principal palavra "páginas" aparecendo no texto escrito.

4.2 Padrão de semelhanças entre conteúdo textual e oral

Alguns pontos da análise tem uma padronização, indepen-

Tabela 6. Ranking de palavras do conteúdo 3

Palavras	Número de aparições textuais	Posição no ranking textual	Número de aparições oral	Posição no ranking oral
Páginas	82	6ª	3	157ª (2ª esfera)
Citações	81	7ª	14	33ª
Universidades	66	11ª	31	10ª
Conexões	55	16ª	6	72ª
UFRGS	51	18ª	24	16ª
UFSC	41	21ª	17	25ª
Websites	39	22ª	17	26ª
Lugar	37	23ª	2	232ª (3ª esfera)
Web	36	25ª	7	69ª (2ª esfera)
Universidade	34	26ª	28	13ª
Produção	16	54ª (2ª esfera)	12	41ª
Página	15	61ª (2ª esfera)	19	20ª
Website	10	97ª (2ª esfera)	16	30ª
Link	6	137ª (2ª esfera)	12	39ª

dente do tipo de conteúdo (textual ou oral), onde de forma geral apresentam um modelo de distribuição normal, em especial pelos experimentos aqui retratados, como a relação das palavras escritas e faladas; a média de aparições destes conteúdos; a média encontrada na primeira esfera, bem como a média da segunda esfera que podem ser canalizadas como palavras triviais ou interessantes, dependendo de como resultam suas aparições e seus conteúdos vazios de significado. Estas vertentes de análises estão expostas no padrão encontrado na figura 4.

5. Conclusões

Os resultados detalhados nos trazem a dimensão do trabalho escrito e sua importância dentro do meio acadêmico e que, de um modo geral, a apresentação oral nem sempre traz os detalhes mais importantes que foram descritos na parte escrita; levando em consideração, principalmente, que no modo escrito não há tantos 'vícios de linguagem' e informalidade como nas apresentações orais; também deve-se analisar os fatores humanos que há em uma apresentação, como a interação com os ouvintes.

> "No modo escrito não há tantos 'vícios de linguagem' e informalidade como nas apresentações orais"

Ao analisar os dados, é possível perceber que o conteúdo textual traz uma base forte para a apresentação oral, onde em média, a cada 2 palavras escritas 1 palavra aparece no conteúdo oral.

Destacamos a seguir algumas considerações finais de todos os textos:

- o texto escrito ainda não pode ser substituído pela apresentação oral;
- as apresentações orais, em sua maioria, buscam palavras diferentes para explicar o que foi feito no trabalho escrito;
- não há uma preocupação ao detalhar o trabalho escrito em uma apresentação oral, uma vez que análise individual das palavras não é feita. ;
- a primeira esfera é maior nas apresentações orais, porém, como analisado, nem sempre esta primeira esfera pode ser levada em consideração como "norteador" do trabalho escrito;
- de modo geral, a segunda esfera no texto oral seria a parte mais importante. Como dito anteriormente, há vícios de linguagem e expressões informais/rotineiras que só fazem sentido acompanhadas do gestual e expressões faciais do apresentador do trabalho que acabam interferindo no sentido principal da primeira esfera e, por isso, estas palavras acabam sendo encontradas apenas na segunda parte. Com base nos conteúdos observados, foi possível encontrar uma "semelhança inversa" entre as palavras ditas e escritas. Em todos eles, os rankings tiveram posições inversas –por exemplo a palavra mais escrita foi uma das menos faladas, e vice-versa.

Em uma ampliação para futuros estudos pode-se utilizar a metodologia proposta para investigações de natureza informétrica, visando quantificar o processo de recuperação, relevância e revocação informacional dos textos, atrelando a uma prática a indexação automatizada, organização e representação da informação. Ainda visionando novos estudos, é necessário entender o porquê do texto oral fugir do contexto geral do trabalho escrito, uma vez que esta apresentação é baseada, totalmente, no trabalho escrito.

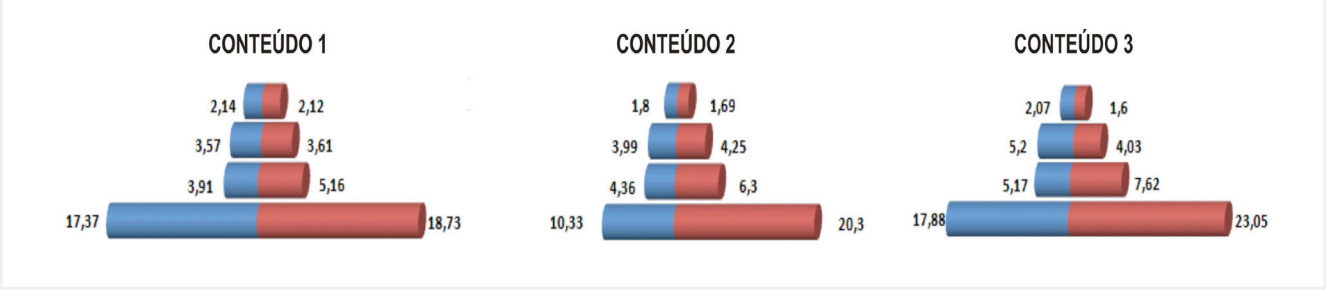
Outra funcionalidade deste tipo de método está voltada a questão de análise do discurso e transcrição de áudio, muito utilizada pelo jornalismo. Em um plano mais ambicioso também pode servir para identificar metodologias utilizadas nos estudos e suas equivalências as áreas de conhecimento.

Em âmbito geral, a aplicação está voltada para os estudos métricos, porém pode ser adaptada para qualquer estudo que tenha como pano de fundo a quantificação de palavras nos textos apenas levando em consideração o objetivo a ser alcançado, uma vez que essa metodologia depende de alguns fatores subjetivos como classificar separadamente ou não substantivos no singular ou plural.

De forma geral, existem padrões de semelhança entre o que é escrito e o que é falado, pois de cada duas palavras escritas uma é falada, tendo uma relevância considerável; a própria média de aparições contém semelhanças, o mesmo acontece quando dividimos por esferas que apresentam percentuais semelhantes.

Finalmente, o estudo se caracterizou em desvendar as semelhanças e diferenças entre a forma escrita e a oral, aplicando uma lei clássica da bibliometria, porém foi difícil conseguir literatura atual da aplicação de Zipf que pudesse ser utilizada no estudo, visto que as inovações desta lei se restringem a analisar a influencia de revistas e de termos para as áreas de conhecimento. Deste modo podemos afirmar que a análise é totalmente inédita e por este motivo a literatura utilizada se restringe a uma vida média intermediária da literatura.

Figura 4. Padrão dos três conteúdos (textual e oral)



6. Referências

Abhari, Abdolreza; **Soraya, Mojgan** (2010). "Workload generation for *YouTube". Multimedia tools and applications*, v. 46, n. 1, pp. 91-118.
http://dx.doi.org/10.1007/s11042-009-0309-5

Baayen, Harald (1992). "Statistical-models for word-frequency distributions: a linguistic evaluation". *Computers and the humanities*, v. 26, n. 5-6, pp. 347-363.
http://dx.doi.org/10.1007/BF00136980

Chen, Ye-Sho; **Chong, Pete** (1992). "Mathematical-modeling of empirical laws in computer-applications: a case-study". *Computers & mathematics with applications*, v. 24, n. 7, pp. 77-87.
http://dx.doi.org/10.1016/0898-1221(92)90156-C

Chen, Ye-Sho; **Leimkuhler, Ferdinand F.** (1987). "Analysis of Zipf's law: an index approach". *Information processing & management*, Great Britain, v. 23, n. 3, pp. 171-182.
http://dx.doi.org/10.1016/0306-4573(87)90002-1

Danesi, Marcel (2009). "Explaining change in language: a cybersemiotic perspective". *Entropy*, v. 11, n. 4, pp. 1055-1072.
http://dx.doi.org/10.3390/e11041055

Egghe, Leo (1991). "The exact place of Zipf's and Pareto's law amongst the classical informetric laws". *Scientometrics*, v. 20, n. 1, pp. 93-106.
http://dx.doi.org/10.1007/BF02018147

Garner, Philip N. (1997). "On topic identification and dialogue move recognition". *Computer speech and language*, v. 11, pp. 275-306.
http://dx.doi.org/10.1006/csla.1997.0032

Günther, Ralf; **Levitin, Lev**; **Schapiro, Boris**; **Wagner, Peter** (1996). "Zipf's law and the effect of ranking on probability distributions". *International journal of theorical physics*, v. 35, n. 2, pp. 395-417. 1996.
http://link.springer.com/article/10.1007/BF02083823

Naranan, Sundaresan; **Balasubrahmanyan, Vriddhachalam K.** (1992). "Information theoretic models in statistical linguistics. Part I: A model for word frequencies". *Current science*, v. 63, n. 5, pp. 261-269.
http://www.currentscience.ac.in/Downloads/article_id_063_05_0261_0269_0.pdf

Ohnishi, Teruaki (1993). "Selective amplification of the amount of nuclear information released by the newsmedia". *Annals of nuclear energy*, v. 20, n. 8, pp. 525-532.
http://dx.doi.org/10.1016/0306-4549(93)90001-6

Paliwal, H. V.; **Bhatnagar, S. N.**; **Haldar, S. K.** (1986). "Lead-zinc resource prediction in India: an application of Zipf's law". *Mathematical geology*, v. 18, n. 6, pp. 539-549.
http://dx.doi.org/10.1007/BF00914254

Quoniam, Luc; **Balme, Frédéric**; **Rostaing, Hervé**; **Giraud, Eric**; **Dou, Jean-Marie** (1998). "Bibliometric law used for information retrieval". *Scientometrics*, v. 41, n. 1-2, pp. 83-91.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.4482&rep=rep1&type=pdf
http://dx.doi.org/10.1007/BF02457969

Ridley, Dennis R. (1982). "Zipf's law in transcribed speech". *Psychological research*, v. 44, pp. 93-103.
http://dx.doi.org/10.1007/BF00308559

Rousseau, Ronald; **Zhang, Qiaoqiao** (1992). "Zipf data on the frequency of Chinese words revisited". *Scientometrics*, v. 24, n. 2, pp. 201-220.
http://dx.doi.org/10.1007/BF02017909

Schaer, Philipp (2013). "Information retrieval und informetrie: zur anwendung informetrischer methoden in digitalen bibliotheken". *Historical social research - Historische sozialforschung*, v. 38, n. 3, pp. 282-354.
http://eprints.rclis.org/22631/1/HSR_38.3_Schaer_b.pdf

Serrano, M. Ángeles; **Flammini, Alessandro**; **Menczer, Filippo** (2009). "Modeling statistical properties of written text". *PLoS one*, v. 4, n. 4, pp. 5372-5376.
http://dx.doi.org/10.1371/journal.pone.0005372

Shan, Shi (2005). "On the generalized Zipf distribution: Part I". *Information processing & management*, v. 41, n. 6, pp. 1369-1386.
http://dx.doi.org/10.1016/j.ipm.2005.03.003

Shtrikman, S. (1994). "Some comments on Zipf law for the Chinese language". *Journal of information science*, v. 20, n. 2, pp. 142-143.
http://dx.doi.org/10.1177/016555159402000208

Wang, Haiyang; **Liu, Jiangchuan**; **Xu, Ke** (2012). "Understand traffic locality of peer-to-peer video file swarming". *Computer communications*, v. 35, n. 15, pp. 1930-1937.
http://dx.doi.org/10.1016/j.comcom.2012.06.003

Yu, Jiang; **Chou, Chun-Tung**; **Yang, ZongKai**; **Du, Xu**; **Wang, Tai** (2006). "A dynamic caching algorithm based on internal popularity distribution of streaming media". *Multimedia systems*, v. 12, n. 2, pp. 135-149.
http://dx.doi.org/10.1007/s00530-006-0045-x

Zhang, ZiKe; **Lv, Linyuan**; **Liu, Jian-Guo**; **Zhou, Tao** (2008). "Empirical analysis on a keyword-based semantic system". *European physical journal B*, v. 66, n. 4, pp. 557-561.
http://dx.doi.org/10.1140/epjb/e2008-00453-9

Zhou, Xiaobo; **Xu, Cheng-Zhong** (2007). "Efficient algorithms of video replication and placement on a cluster of streaming servers". *Journal of network and computer applications*, v. 30, n. 2, pp. 515-540.
http://dx.doi.org/10.1016/j.jnca.2006.03.001

Zipf, George-K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley: Cambridge Mass, 543 pp.

Zörnig, Peter; **Altmann, Gabriel** (1995). "Unified representation of Zipf distributions". *Computational statistics & data analysis*, v. 19, n. 4, pp. 461-473.
http://dx.doi.org/10.1016/0167-9473(94)00009-8